Leszek Bajkowski
Uniwersytet Jagielloński
Kraków

Before they can teach they must talk: On some aspects of human-computer interaction

Introduction

While promising technological advances have been made in the areas of speech recognition, generation and understanding, developing usable dialogue systems is still difficult as researchers find themselves in need for models of spoken discourse. The models should cover such phenomena as entrainment, turn-taking or dynamic adaptation by providing insight into both human-human and human-machine interaction and their similarities and differences. The work in the field of human-computer interaction (HCI) aids the creation of human-machine interfaces, allowing for spoken communication which is as close as possible to human-human interaction in natural language. The systems developed are designed to assist people with little or no technological insight: a doctor in his medical decisions, a tourist in a foreign city, a customer taken through a display of products, or a machine operator at his place of work. They are also designed to aid a learner in their attempts at mastering a foreign language by way of interaction with a virtual agent. Besides natural language understanding and generation, such agents are being endowed today with other human-like features in an attempt to make them resemble communication between humans as closely as possible. If systems like these are to be viable, they must be able to process spoken language in its many layers – phonetic, semantic and pragmatic – and, at the same time, perform some useful tasks like giving advice, providing information or teaching a language. Each of these inspire a separate field of research in which technical and theoretical problems are identified and attempts to overcome them undertaken.

Chatbots

In the early twentieth century the English mathematician Alan Turing asked the question "Can a machine think?" and devised a test to verify if it really could. If a machine gives an impression that it can interact with a human being in an intelligent way, it has

passed the test. It is not a coincidence that the test is based on dialogue. If a human judge who interacts with a machine thinks he is talking to another human, then the machine may be said to think. The famous Turing Test is taken by numerous conversational programs entered for a yearly Loebner Prize competition (http: //www.loebner.net/). A New York philanthropist Hugh Loebner offers money each year ($2250 in 2007) to the programmer whose program called a "chatting robot" (chatbot or chatterbot for short) best imitates human responses. The first chatbot, a program published in 1966 by an MIT professor Joseph Weizenbaum, was called Eliza. Eliza and similar programs imitate communication through a text-based dialogue. They make use of the text input from the interlocutor by applying some simple pattern-matching techniques. The algorithm tries to find keywords in the input and select a response out of its store of responses. As a result if you tell the chatbot "I need your advice", it will respond with "How badly do you need it?" and if you say "I need a new pair of shoes", the answer will be exactly the same. Chatbots are innocent of true syntactic parsing, meaning, pragmatics, and any other linguistic knowledge.

Granted that chatbots have a limited application, they are not completely useless from the point of view of computer-assisted language instruction. First, there is no doubt that even limited exchanges, i.e. dialogues that sometimes go off a cohesive course, can be fun for learners and allow some form of language practice. Second, let us notice that text-based chatbots require the user to enter well-spelled word-forms and, actually, the simpler the text entered (i.e. the more limited the structure and the semantic content of the text), the better chance there is for the dialogue to carry on smoothly. Third, it is possible to download a package of chatbot-building software and even some ready-made content (e.g. some factual statements which will stand for the bot's "knowledge") and write one's own language teaching chatbots using AIML language (some components can be downloaded from www.alicebot.org/downloads/). Also, the teacher can simply use one of the existing on-line bots (a good point to start online is the Simon Laven Page at www.simonlaven.com/). Dave E.S.L., an English-teaching chatbot (www.alicebot.org/dave.html), already exists but requires paid subscription.

It must be remembered that chatting with a simple Loebner Prize contestant is keyboard-based and there are differences between spoken and written language: the inner structure of a spoken dialogue is not the same as the inner structure of a written communication; in spoken dialogue less time is given to language user to formulate the utterance; speech production errors may be corrected; instantaneous error correction and the possibility of making an error introduces hesitation; speech is more broken and looser than writing; there is also some redundant information, repetitions, restarts, interjections, contradictions, and even a tendency to stop the interlocutor (Minker and Bennacef, 2004). Such disfluencies and self-corrections account for 6–15% of what is said. Spoken dialogue is best conceptualized as a cooperative activity to which both partners continually contribute on a continuum from the purely verbal to the entirely non-verbal, involving behavior at three levels: the level of words, the paralinguistic level, and – in some cases – the visual level (Schober, 2006).

Automatic speech recognition

Spoken communication is the most natural and a very efficient way of transmitting language: an average person can type in 20 words on a keyboard, write 24 words on paper, and utter some 150 words per minute (Minker and Bennacef, 2004). There is always an advantage in using spoken language in the accomplishment of a task, especially if compared with using a keyboard. But before systems can converse with humans, the system must be able to understand what is said and it must be able to speak itself. There is no doubt that the research in the areas of automatic speech recognition (ASR) and speech synthesis are well-motivated. In both areas engineers face some serious technical problems which must be resolved before HCI truly resembles a natural dialogue between humans.

Already in 1950s and 60s serious work started in the area of ASR and the second half of the twentieth century marks its steady progress. Researchers from several American, English and Japanese laboratories presented their first recognizers. In 1970s the first systems capable of recognizing human speech were developed and commercialized. Since then there has been a transition from simple dialogue systems to almost fully conversational. Today's speech technologies are commercially available for a limited range of tasks, enabling machines to respond correctly and reliably to human voices, and provide useful and valuable services (Juang and Rabiner, 2006).

ASR technology is the process of converting incoming phone strings into a sequence of known words. Discrete speech recognizers require that the speaker use discrete speech by introducing pauses of about 200 msec in between the words. This is an unnatural manner of speaking, tedious and difficult to learn but with respect to CALL the solution may be preferred for some applications where practice in the pronunciation of separate words or sounds is the primary goal. With respect to continuous speech, two major approaches to ASR can be well represented by the efforts of IBM and AT&T Bell Laboratories leading, respectively, to creating two types of ASR systems: speaker-dependent and speaker-independent. Since IBM wanted to develop a voice-activated typewriter performing transcription, they developed a speaker-dependent tool that needs to be trained; namely, the user must spend some time reading words and short texts to teach the system about his voice characteristics. A user-trained system can generally cover a large vocabulary at normal pace with a high accuracy. On the other hand, AT&T was after an automated telecommunication public service expected to work well for millions of talkers. Speech recognition algorithms were based on an acoustic model (spectral representation of sounds or words). The acoustic model is trained on a corpus of audio samples. The audio recordings are transcribed and then a statistically represented as speech models. The result is a speaker-independent system which can recognize a smaller number of words but accept input from many different speakers, often with notably different accents. Speech recognition research in the 1980s was characterized by a shift in methodology from the more intuitive template-based approach (templates were recorded patterns to which input was matched) toward a more rigorous statistical modeling framework. Thus while technology was realized differently in various applications, the statistical methods caused a certain degree of convergence in the system design. Most notably the method known as Hidden Markov Model, which

measures the probability of the next element in a string, has been wildly implemented (Juang and Rabiner, 2006).

Accuracy of ASR systems has always been constrained by practical factors such as background noise, accent, sound card quality, pre-amplifying, microphone type and placement, fatigue and stress of users induced by the task or type of utterance. Various methods were tried in response to the problem of corrupted input. In one method, called keyword spotting, the presence of a key-phrase is sufficient to indicate the caller's intent so that the system can trigger an appropriate response. This simple substitute of semantic understanding generally works well because the vocabulary used in the context is limited and therefore less ambiguous than in free dialogue. Systems that employ ASR are designed to allow humans to converse with a database containing information about some specific domain. They are also geared towards performing specific tasks (e.g. providing information about hotels or tutoring about physics) and the HCI takes place in a specific context where the users are expected to conform to linguistic conventions (in a way, there is a limited set of ways to, say, ask about the train times). Therefore the ASR systems are mostly task-oriented and domain-specific.

From the perspective of CALL the most important consideration is how effective the ASR techniques actually are. The performance of ASR systems is specified in terms of accuracy (word error rate) and speed, measured with the real time factor, i.e. how much time is needed to process a recording of a certain duration. Manufacturers generally provide prospective buyers with very high accuracy figures in the range between 95–99%. With respect to the early systems, the claim was verified in a series of laboratory experiments and the researchers reported it was nearer 70%. The answer is in fact always technology-dependent and always related to the human-machine interface design. Also, according to Lamel and Gauvain (2003), tasks should be divided according to which error rates are measured. In small vocabulary tasks (isolated common words) the error rate is below 1%; for read speech tasks (approximately 1000 words) the rate is 3%; and for large vocabulary tasks word-error rates around 8% were obtained using a 65k vocabulary (the data come from 1995, the texts read to the system were newspaper articles from the *Wall Street Journal*). The tasks are different and yield a wide range of results. It is difficult to compare results across the systems; besides, benchmarks for many of the contemporary systems are not publicly available. The best accuracy reported in 1990s with respect to "raw" recognition of continuous speech (e.g. recognition of texts from the *New York Times* read by humans) was 95%, i.e. one recognition error in every twenty words (Ehsani and Knodt, 1998).

Automatic speech generation

Parallel to ASR, speech generation has developed driven by the same technological developments. Today the technology is mainly used in the task of text-to-speech synthesis (TTS), i.e. converting text (of unrestricted vocabulary) into intelligible speech, but building machines that can imitate human ability to talk has a long history that starts with purely mechanical devices. In the eighteenth century inventors like Kretzenstein, Von Kempelen

or Wheatstone tried to build acoustic speaking machines. The technological milestone of the first half of the twentieth century was Homer Dudley's speech synthesizer called the VODER, which was demonstrated in New York in 1939. In 1961 Bell Laboratories showed a computer-based speech synthesis system, a demonstration seen by the author Arthur C. Clarke, giving him the inspiration for the talking computer HAL in "2001: A Space Odyssey". Application potential of synthetic speech is enormous. Speech synthesis provides important utility in situations where a person's eyes and hands are busy, they are used extensively by telecommunication industry or in devices that read out loud for the blind, in video games or children's toys or provide voice for sufferers of neurological disorders: for example, the physicist Stephen Hawking used DECtalk developed by Dennis Klatt.

There are three methods used to generate synthetic speech: concatenative, formant, and articulatory synthesis. Most modern commercial TTS systems are based on concatenative synthesis, in which samples of speech are chopped up, stored in a database, and combined and reconfigured to create new sentences. In terms of its quality, today's synthetic speech is nearing the level of natural speech but it is still not equivalent to it. Its perception and comprehension are still worse than natural speech, and it degrades faster under adverse listening conditions (e.g., noise, distraction, divided attention). Besides, when tasks are demanding, the ability to understand synthetic speech is adversely affected (for example, pilots reported not hearing any synthesized warning sounds or messages when they were in a dangerous situation). Conversely, the more experience a listener has with a particular text-to-speech system, the higher the intelligibility of the speech (for example, blind programmers who routinely used a low-intelligibility synthesizer called Votrax Type-'n-Talk reported that the speech quality improved substantially). Some improvements in the field have come from the work with prosodic rules (like stress or pitch contour) which adjust pronunciations to contexts. Other solutions could be provided by work on affective features, a technology known as emotional speech. By adding emotional quality to synthesized voice, affective information is added and, as a consequence, unnaturalness reduced. Emotional speech synthesis is also important for the use of synthetic speech as prosthesis for voice-disabled individuals that need to rely on synthetic speech for conveying the same range of information conveyed by human speech (Jurafsky and Martin, 2006; Nusbaum and Shintel, 2006).

Talking heads and pedagogical agents

Many of today's instructional tutoring systems incorporate both the speech recognition and generation components, in which case the latter carries the auditory model of the foreign language. This can be significantly enhanced by a variety of visual aids. A very good example of a successful technology using visual enhancement is Baldi, a part of CSLU Speech Toolkit (http: //www.cslu.edu/toolkit/) – an authoring environment for building spoken language systems. It includes recognition modules, text-to-speech synthesis and facial animation in the form of a 3-D "talking head", which provides realistic speech organs movements synchronized with the audio. The user of this software can see the speech organs and how they move to produce a particular sound. The animation

technology allows to visualize processes giving the kind of insight that is not possible even with a living human being. The motivation behind this technology is the audio-visual speech studies reinvigorated by McGurk and MacDonald (1976). They demonstrated that seeing the face of the speaker influences the way we hear the sound. The creators of Baldi consider speech not just an auditory but a multimodal phenomenon. Perception and understanding are not only enhanced by a speaker's face and gestures: visual and auditory speech are complementary (one of the sources is strong when the other is weak) and there is an optimal integration of the two sources of information. The animated face can be aligned with the output of a speech synthesizer or natural auditory speech. It also allows synthesis of several other languages except English, such as Spanish, Arabic, Chinese, German, and Russian. Other studies also demonstrated the influence of vision on audition. For example, it has been estimated that seeing the talker offers a gain equivalent to around a 15 dB increase in signal-to-noise ratio for users of English. What the face talking to you looks like also seems to have an effect on speech perception: talkers who appear to be Japanese generate different effects than those who appear to be American (Campbell, 2006; Massaro *et al.*, 2006).

"Talking heads" have been used as language tutors for both native and second-language learners and individuals with special needs. In experiments evaluating the effectiveness of the tool, the fact exploited most was the multiplicity of sources of information in perception, recognition, learning, and retention. Researchers put emphasis on the fact that words are experienced from many different perspectives: the learner observes them being spoken by a realistic-looking talking head, sees them in writing, sees visual images of referents of the words, types them, hears him- or herself say the words, and compares them with the correct pronunciation of words used in context. The experiments showed that the application could be used successfully with reading-impaired and autistic children, children with hearing loss, and regular English language learners alike to teach new vocabulary (Massaro *et al.*, 2006).

Visualization embedded in the system interface allows for a two-way multi-modal contact between the user and the system. Unlike direct manipulation interfaces, intelligent, interactive, animated pedagogical agents offer an ease of access to the system (Brennan, 1998). Since they take the shape of a 2-D or 3-D character and are often placed in a virtual environment, they can be seen and addressed. They can demonstrate principles, procedures and actions; they can use gestures, gaze and other visual aids to teach but also to attract the student's attention or regulate turn-taking in a mixed-initiative dialogue. Head nods and facial expressions, which are natural devices present in a human dialogue, can provide unobtrusive feedback on the student's utterances and actions. The mere presence of a lifelike agent may increase the student's motivation to perform the task well. While animated pedagogical agents increase the computer's ability to engage and motivate students, there are also quite a lot of things they should be able to do. They need to give the user an impression of being knowledgeable, attentive, helpful, or concerned. In short, they should look believable enough to support pedagogical interactions. Besides, pedagogical agents must exhibit flexibility in order to manage the learning environment and the student, with their unpredictable aptitudes, levels of proficiency, and learning styles. All this means that the implemented model must be dynamic and adaptive, as opposed to deliberate, sequential, or preplanned.

Managing human-machine dialogue

Of course to recognize words and sentences and generate human-like voice is one thing; making a conversation is another. Speech applications which support spontaneous and conversational styles often require a dialogue between the user and the machine to reach some desired state of understanding. Such a dialogue often requires such operations as query and confirmation, thus making some allowance for errors and repairs. These factors focused the attention of the research community on the dialogue management component. MIT's Pegasus and Jupiter and the HMIHY (which stands for How May I Help You?) developed at AT&T are noteworthy for their ability to manage dialogue effectively. Pegasus provides information about the status of airline flights; Jupiter can be asked about the weather conditions (Juang and Rabiner, 2006).

Dialogue management seems to be the key to successful virtual tutoring. A convincing dialogue system must be able to maintain a pragmatically appropriate conversation which includes the ability to handle semantic trickiness (e.g. anaphors, ellipses or synonymy), identify the topic, use backchannels, know when to take turns, or establish common ground. In CALL settings it should also have some sort of mechanism to control the task completion and provide feedback. Some effort towards using dialogue systems for language instruction has already been made. After adapting a telephone-based bus schedule information system called Let's Go for CALL (for speech recognition and correction prompt generation) Raux and Eskenazi (2004) conclude that it can be used to provide realistic, involved environment for language learning. They say there is potential for CALL in those systems as they place the student in a realistic situation where a specific task has to be accomplished in the target language. Unfortunately, such systems assume that the users' language is perfect and any disruption is due to speech recognition errors on the part of the machine and therefore some acoustic and lexical improvements have to be made.

Today's language teaching systems, for example Herr Kommissar, Subarashii, Virtual Conversations or TLTS, analyze pronunciation or teach speaking through games or role play (Johnson *et al.*, 2004), but they do not support a fully complex open-ended dialogue. Subarashii and Steve are good examples of existing virtual tutors. Subarashii is a language tutor used to teach beginner's Japanese through virtual spoken interactions by role play (like inviting a friend to the cinema). An animated agent takes one of the roles in the dialogue. It is designed to understand what a student says and respond in a meaningful way in spoken Japanese. Although the student is not directly presented with a list of response choices, the exercises are constrained because the expected responses are still limited to a few correct ones. Despite the restricted communicative competence of second language beginners, the range of utterances, both valid and invalid, which they can produce is huge. Subarashii successfully processes all correct utterances but also recognizes and rejects many incorrect ones. Experiments with students in California have shown that "ASR works in a school setting, and that high school students seem to enjoy it" (Bernstein *et al.*, 1999).

While the Subarashii interlocutors are 2-D animated persons, Steve (Soar Training Expert for Virtual Environments) can be seen by the student in stereoscopic 3-D. The agent can also see the student thanks to the virtual environment's tracking hardware

which monitors the student's position and orientation in the environment. More importantly, however, it is designed to interact within the virtual environment. Steve can adapt his demonstrations in midstream if the student performs actions that interact with the demonstration and respond to student interruptions. Steve has been applied to naval training tasks such as operating the engines aboard US Navy ships. The authors say that the immaturity of the technology prevents any comprehensive, definite empirical studies of the effectiveness of animated pedagogical agents. However, they quote a study conducted with one hundred middle school students the purpose of which was to obtain a "baseline" reading of the effectiveness and impact of various forms of agent advice. The primary conclusion was that students interacting with learning environments with an animated pedagogical agent show "statistically significant increases from pre-tests to post-tests" (Johnson *et al.*, 2000).

Systems like Let's Go, Subarashii or Steve cannot handle a completely unconstrained spoken conversation, and other systems which attempt to do so generally fail (Jurafsky and Martin, 2000). The issue, as it seems, is the lack of applicable models of interaction. These models are being sought and dialogue is being analyzed with respect to both human-human and human-machine interaction. Because computer talk (another label for HCI, by analogy to "baby talk") reveals some features that are replicated and observable again and again, the conclusion is that there exists a regular register for humans conversing with dialogue systems. Here are some reasons why HCI is not a simple replication of regular dialogue: in communication with the machine, apart from its manifestly task-oriented and domain-specific character, people display a range of adaptive language behaviors (Minker and Bennacef, 2004; Leech and Weisser, 2003). The adaptation is similar to that found in human-human communication but it also has its peculiarities.

Cooperation through alignment

One way people change their behavior in the course of a dialogue – alignment – has been intensively studied in the context of HCI. In the cooperation process people try to align with the conversation through back-channeling (responses like "yes", "hmmm", "I see", "uh-huh", facial expressions, eye-contact, nods and gestures), common ground (sharing knowledge, sentiments, associations), and some social factors (such as talking to a child or a superior). The tendency to align is most vivid in entrainment, a cooperative reuse of each other's language. Garrod and Anderson (1987 cited in Porzel, 2006) suggest that people tend to automatically converge on lexical and syntactic choices via a low-level mechanism of interpersonal priming. The hypothesis is that the participants attempt to align their situation models, i.e. their common understanding of the situation. At a lexical level alignment is achieved through the choice of words out of the vocabulary used by the interlocutor. The one who introduces a term is denoted as the leader and the one who adopts it as the follower. First, the interlocutors need to establish common ground for their conversation. After that they hedge, i.e. they mark the term as provisional, pending evidence of acceptance from the other. Only then do they agree on the same choice of words. As a last step, the terms are no longer indefinite and can

be shortened, e.g. via anaphora, pronominalization or gapping (Porzel, 2006). Brennan (1996), who explored conceptual pacts people make during conversation, says that the shared conceptualization between interlocutors is marked by using the same terms to refer to the same objects. While the likelihood that people in one conversation would choose the same terms for the same common object as people in another conversation is only 10%, in a situation when two people repeatedly discuss the same object, they come to use the same terms and stay with them for the rest of the dialogue. Brennan (1996) has also shown that people will align their language towards that of computer agents.

In the case of HCI, the users' lexical choice seems to be strongly influenced by their beliefs and expectations about the system. Pearson *et al.* (2006) report that when users believe that the system is unsophisticated and restricted in capability, they adapt their language to match the system's language more than when they believe the system is relatively sophisticated and capable. Interestingly, the actual behavior of the system is irrelevant. Bhatt *et al.* (2004) studied student use of hedges (expressions like "I guess", "maybe", "kind of", "I'm not sure", etc.) and affect (present in expressions like "sorry", "wow", "I get it", "I'm a bit confused") in interacting with both humans and computer systems, during keyboard-mediated natural language tutoring sessions in medicine. They say that while hedging always occurred in human-human dialogues, there was virtually no hedging when subjects addressed machines. As opposed to hedging, students do express affect to machines, though far less often than to humans. They do not apologize or thank or give computers direct feedback; they do, however, express confusion, frustration and even rage. The conclusion is that students who know they are interacting with a computer change their attitude towards the conversation, and they are less concerned with helping to keep the flow going than they are in regular conversation. Other authors have come to similar conclusions: in a dialogue with a machine, humans change their attitude. But they may act reasonably if prompted to do so. This is illustrated by the issue of turn-taking behavior of "naive users", i.e. persons who communicate with the system without any or with too little prior experience to adjust linguistic behavior. These are situations where the human interlocutor makes an additional or second utterance before the system has provided its response to the first utterance. After that the conversation becomes asynchronous; the system responds to the last but one utterance while in the user's mind that response concerns the last. In consequence, the dialogue fails completely. Clearly, conversational dialogue systems suffer from a lack of strategies that would prevent the dialogue from becoming "out of sync" and allow repairs (Porzel, 2006). However, this may be remedied by the dialogue management component which should keep the user informed about the process and provide messages telling him to stand by or restart (e.g. "I listen to you, please go on").

Conclusion

There is no denying that technology has made huge advances in the recent years in such areas as speech synthesis and recognition, visualization, and pragmatically motivated dialogue control. The hope is that many of the partial capacities of various state-of-the-

art systems will be eventually put together to make a system that can actually engage in a sensible conversation and teach a language in a technologically unconstrained fashion. A review of literature and online resources shows that there are still many features of human speech that are problematic and machines passing the Turing Test are not imminent, but as technology continues to mature, the number of valuable systems that become part of our daily life is growing. At the same time some frustration is visible concerning the missing theoretical background in terms of a clear account of the nature of human-human, human-computer (and recently also computer-human) interaction that would allow systems designers to look for engineering solutions. As for CALL-oriented applications, there is always a need for empirically verified knowledge about their effectiveness in real learning context. Some grounds, however, exist to believe that, once useful interaction models are available and technical hurdles disappear, teachers and learners may obtain effective and attractive technological support to complement their human effort in the area of language teaching.

Bibliography

Bernstein, J., A. Najmi, F. Ehsani, 1999, "Subarashii: encounters in Japanese spoken language education", *CALICO Journal* 16 (3): 361–384. Accessed through <www.calico.org/html/article_619.pdf> (10 August 2007).

Bhatt, K., M. Evans, S. Argamon, 2004, "Hedged responses and expressions of affect in human/human and human/computer tutorial interactions". Accessed through <lingcog.iit.edu/doc/bhattevansargamonsubmit.pdf> (10 August 2007).

Brennan, S.E., 1996, "Lexical entrainment in spontaneous dialog", *International Symposium on Spoken Dialog ISSD-96*, 41–44. Accessed through <www.psychology.sunysb.edu/ sbrennan-/papers/brenissd.pdf> (10 August 2007).

Brennan, S.E., 1998, "The grounding problem in conversations with and through computers", *Social and Cognitive Psychological Approaches to Interpersonal Communication*, 201–225, Hillsdale, NJ: Lawrence Erlbaum. Accessed through <www.psychology.sunysb.edu/sbrennan-/papers/brenfuss.pdf > (10 August 2007).

Campbell, R., 2006, "Audio-visual speech processing", [in:] K. Brown, ed., *Elsevier Encyclopedia of Language and Linguistics*, Oxford: Elsevier Ltd, pp. 562–569. Accessed through <www.uj.edu.pl> (2 March 2007).

Ehsani, F., E. Knodt, 1998, "Speech technology in computer-aided language learning: strengths and limitations of a new CALL paradigm", *Language Learning and Technology* 2(1): 45–60. Accessed through <llt.msu.edu/vol2num1/article3/> (10 March 2007).

Garrod S., A. Anderson., 1987, "Saying what you mean in dialogue: A study in conceptual and semantic coordination", *Cognition* 27: 181–218.

Johnson, W.L., J.W. Rickel, J.C. Lester, 2000, "Animated pedagogical agents: face-to-face interaction in interactive learning environments", *International Journal of Artificial Intelligence in Education* 11: 47–78. Accessed through <www.isi.edu/isd/carte> (12 August 2007).

Johnson, L., S. Choi., S. Marsella, N. Mote, S. Narayanan, H. Vilhjlmsson, S. Wu, 2004, "Tactical language training system: supporting the rapid acquisition of foreign language and cultural skills", *Proceedings of InStil*, July 2004. Accessed through <sail.usc.edu/publications/TLTS-InSTIL-General.pdf> (15 August 2007).

Juang, B.H., L.R. Rabiner, 2006, "Automatic Speech Recognition-A Brief History of the Technology", [in:] K. Brown, ed., *Elsevier Encyclopedia of Language and Linguistics*, Oxford: Elsevier Ltd, pp. 806–819. Accessed through <www.uj.edu.pl> (2 March 2007).

Jurafsky, D., J.H. Martin, 2000, *Speech and Language Processing: An Introduction to Natural Language Processing*, *Computational Linguistics*, *and Speech Recognition*, New Jersey: Prentice Hall. Accessed through <www.cs.colorado.edu/~martin/SLP/slp.html> (10 June 2007).

Lamel, L., L. Gauvain, 2003, "Speech Recognition", [in:] R. Mitkov, ed., Oxford Handbook of Computational Linguistics, Oxford: Oxford University Press, pp. 305–322.

Leech, G., M. Weisser, 2003, "Pragmatics and Dialogue", [in:] R. Mitkov, ed., *Oxford Handbook of Computational Linguistics*, Oxford: Oxford University Press, pp. 136–156.

Massaro, D.W., Y. Liu, T.H. Chen, C.A. Perfetti, 2006, "A multilingual embodied conversational agent for tutoring speech and language learning", *Proceedings of the Ninth International Conference on Spoken Language Processing*, Bonn: Universität Bonn, pp. 825–828. Accessed through <mambo.ucsc.edu/pdf/massarobox.pdf> (28 July 2007).

McGurk, H., J. MacDonald, 1976, "Hearing lips and seeing voices", *Nature*, vol. 264 (5588), 746–748.

Minker, W., S. Bennacef, 2004, *Speech and Human-Machine Dialog*, Boston: Kluwer Academic Publishers.

Nusbaum, H.C., H. Shintel, 2006, "Speech Synthesis", [in:] K. Brown, ed., *Elsevier Encyclopedia of Language and Linguistics*, vol. 12, Oxford: Elsevier Ltd, pp. 19–30. Accessed through <www.uj.edu.pl> (2 March 2007).

Pearson, J., Hu Jiang, H.P. Branigan, M.J. Pickering, C.I. Nass, 2006, "Adaptive language behavior in HCI: how expectations and beliefs about a system affect users' word choice", *Conference on Human Factors in Computing Systems*, pp. 1177–1180. Accessed through <portal.acm.org> (10 June 2007).

Porzel, R., 2006, "How computers (should) talk to humans". *How People Talk to Computers*, *Robots*, *and other Artificial Communication Partners. Proceedings of the Workshop Hansewissenschaftskolleg*, April 21–23. Accessed through <nats-www.informatik.uni-hamburg.de/~fischer/hriproc.pdf> (10 June 2007).

Raux, A., M. Eskenazi, 2004, "Using Task-Oriented Spoken Dialogue Systems for Language Learning: Potential, Practical Applications and Challenges", *InSTIL* 2004, Venice, Italy. Accessed through <www.cs.cmu.edu/~antoine/papers/rauxeskenazi-instil-04.pdf> (12 June 2007).

Schober, M.F., 2006, "Dialogue and Interaction", [in:] K. Brown, ed., *Elsevier Encyclopedia of Language and Linguistics*, Oxford: Elsevier Ltd, pp. 562–571. Accessed through <www.uj.edu.pl> (2 March 2007).